Statistical Modeling of Unemployment Dynamics Using Panel Data

Team 1: Laiyong Li (ll2233), Simeng Li (sl3245), Yiyao Wang (yw2743),

Zhizhao Wang (zw773), Xiaole Yu (xy525)

Cornell University

Project Advisor: Professor David Ruppert

Client: T. Rowe Price

May 12, 2025

1 Executive Summary

This capstone report presents a statistical framework for modeling individual unemployment dynamics using longitudinal panel data from the Panel Study of Income Dynamics (PSID). The project was developed in collaboration with T. Rowe Price's Multi-Asset Division, with the goal of supporting retirement planning strategies through improved understanding of employment risk.

We build two logistic regression models to estimate the probability of employment over the working life cycle. The first model predicts initial employment status at age 24 based

on demographic and economic variables such as education, gender, and labor income. The second model captures year-to-year employment transitions.

Next, we simulate 10,000 individuals whose covariates are sampled to match real-world distributions by Monte Carlo methods. Simulation results reveal both typical labor force participation trends and critical heterogeneity across individuals.

The models highlight meaningful dynamics in employment status and lay the groundwork for future extensions, such as incorporating duration dependence, macro-shocks, and policy interventions.

2 Background

The Lifecycle research team of T. Rowe Price's Multi-Asset Division works on setting strategic allocation of T. Rowe Price target date funds and development of retirement income and personalized lifecycle investment solutions. A huge part of the research for the team will be focusing on demographics and behavior of individual investors in their roles as employees and retirees. Models of investor's earnings are a crucial part of simulation engine to give customized investment solutions. Therefore modeling the unemployment status will be crucial regarding giving solutions to individual investors. Also, the estimation of unemployment spells is an important topic in understanding the labor force to give investors better investment solutions.

3 Objectives

This analysis has three main objectives. First, it aims to identify the key predictors of employment status, helping to understand which factors most influence whether individuals are employed. Second, it involves simulating unemployment dynamics over the lifecycle to capture how individuals' employment status changes throughout their careers. Third, it seeks to analyze simulated unemployment patterns across the life course to discover trends and behaviors. To achieve the last goal, several key unemployment metrics are analyzed, including unemployment rates by age, which track the probability of being unemployed at different career stages; durations of unemployment spells, which measure how frequently individuals experience unemployment; and transition probabilities between employment states, which provide insights into overall labor market dynamics.

4 Data

4.1 Data Description

This project utilized the Panel Study of Income Dynamics (PSID) Panel Study of Income Dynamics 2025, which is the longest-running panel dataset in the US, maintained by the University of Michigan (PSID Data Center). The PSID tracks both individuals and households over time, providing comprehensive and well-documented data on over 50,000 variables including demographic variables, economic variables, education variables, health variables, and housing variables. The dataset is publicly available through the Survey Research Center at the University of Michigan, accessible at the PSID Data Center. Moreover, the dataset

is longitudinal, collected annually until 1997 and biennially thereafter.

Upon our analysis, we specifically used individual and family-level data from 2011 onward to ensure that our research reflects more recent trends and economic conditions. For our dataset, the following variables serve as key predictors in our further analysis of unemployment dynamics.

4.1.1 Demographic Variables

- PersonID: A unique identifier assigned to each individual in the dataset.
- Year: The survey year corresponding to the recorded data.
- Age: The age of the individual in the given survey year.
- Sex of Individual: The recorded gender of the respondent.
- Years Completed Education: The total number of years of formal education completed by the individual.

4.1.2 Employment and Economic Variables

- Labor Income of Head: Total income from wages, salaries, and other labor-related earnings.
- Employment Status: Indicates whether the individual was employed, unemployed, or out of the labor force.

4.1.3 Health and Insurance Variables

- Whether Covered by Insurance Now: Indicates whether the individual is currently covered by any form of health insurance.

4.1.4 Marriage and Family Variables

- Last Known Marital Status: The most recent recorded marital status of the individual (e.g., single, married, divorced, widowed).

4.2 Data Cleaning

After identifying the variable set, we merged data from multiple years into a single analysis dataset, resulting in a dataset containing 36,371 unique individuals and a total of 218,226 observations. To track individuals across different years (2011–2021), we assigned each person a unique computed ID by:

$$\text{ID} = 1000 \times \texttt{ER30001} + \texttt{ER30002}$$

where `ER30001` is the 1968 Family Interview Number and `ER30002` is the Person Number. Additionally, as per the client's requirements, we transformed the dataset from a wide format to a long format to facilitate longitudinal analysis.

One of the biggest problems with the PSID dataset is dealing with missing values. We imputed the missing values for ages for each individual on the basis of the interview time and their previous ages available. We achieved this by first assigning predefined year offsets for each year (e.g., 2011 $\rightarrow$ 0, 2013 $\rightarrow$ 2, ... , 2021 $\rightarrow$ 10). Then we identified each individual's

earliest non-missing age, corresponding year offset, and current year offset, named them respectively as base age, base year offset, and current year offset. After that, we calculated the current age as the base age plus the difference in year offsets.

To prepare the dataset for analysis, a series of variable transformations and encoding procedures were implemented to ensure consistency and interpretability. The variable `Gender` was recoded into a binary indicator, where 0 denotes female and 1 denotes male. The original `Insurance` variable was simplified by excluding non-informative values. The remaining values were recoded such that 1 indicates possession of any form of insurance and 0 indicates none. Similarly, the `Employment Status` variable was transformed into a binary format, with 1 indicating the individual was employed at the time of the survey and 0 indicating temporary unemployment. Educational attainment, originally measured as years of completed schooling, was categorized into three tiers: basic education (1–8 years), secondary education (9–12 years), and higher education (13–17 years). Labor income was inflation-adjusted using the Consumer Price Index (CPI), with 2011 serving as the base year .

The dataset was further refined by restricting the analysis to the household head, as indicated by specific values in the Relation to Head variable, and by retaining only adults between 18 and 70 years. Another critical step in the process was incorporating historical employment information by creating a lag variable. After sorting the dataset by Person ID and Year, a self-join was performed using a two-year shift in the Year variable to attach the lagged employment status to the current records. Records lacking this information were removed, ensuring that every entry in the final dataset had a valid historical employment indicator.

The final cleaned dataset used for model training contained 12,061 unique individuals

and a total of 41,804 observations.

4.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is to discover relationships in the PSID data from 2011 to 2021. Key variables such as age, sex, insurance, education level, and labor income are examined in relation to employment status. This preliminary process guides the selection of the choice of important variables for subsequent modeling.

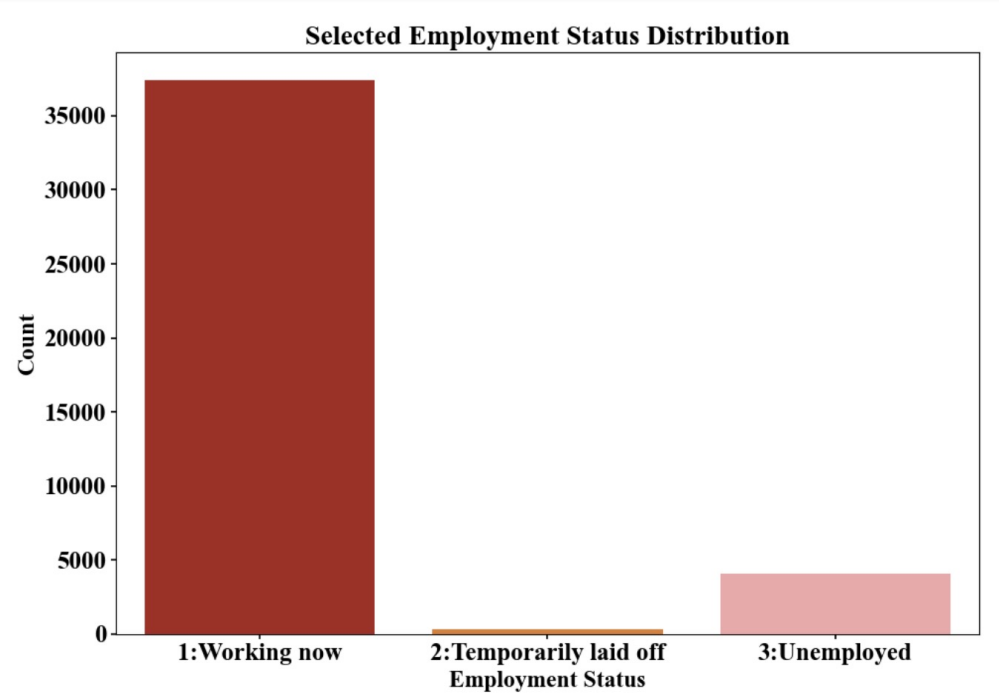

Figure 1: Employment Status Histogram

This is the distribution of employment status. This distribution in Figure 1 suggests a strong labor market participation, with most respondents employed and only a small portion facing temporary or long-term job loss.

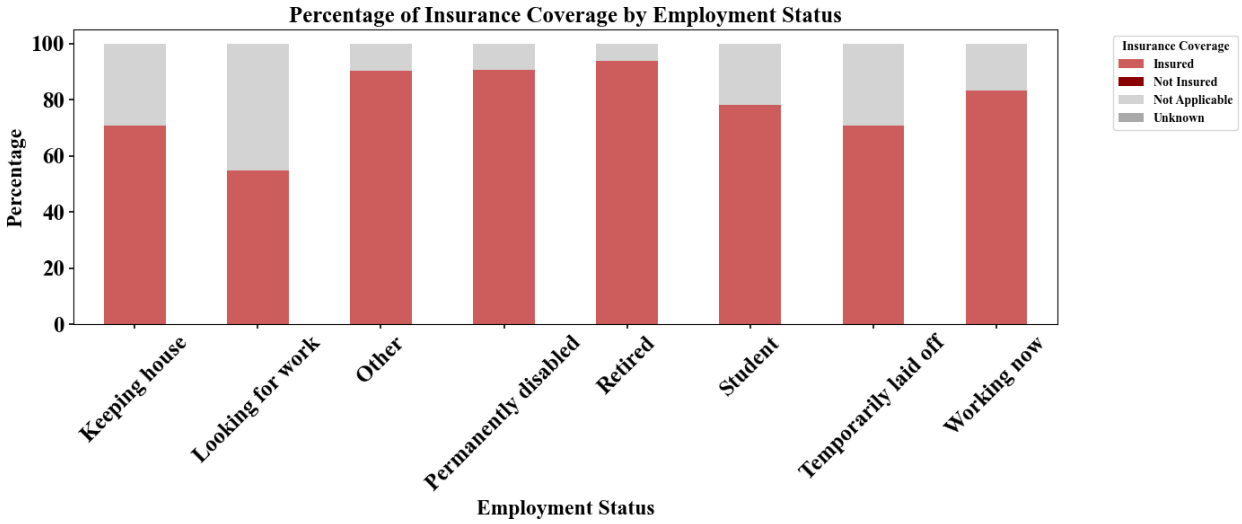

Figure 2: Insurance Coverage by Employment Status

Figure 2 is the percentage-based stacked bar chart of insurance versus employment status. It is obvious that most people are insured especially in the group of "working now", "retired", and "permanently disabled".

Figure 3: Median Labor Income by Employment Status

Figure 3 is the bar chart of labor income versus employment status, "Working now" clearly has the highest median income, and other groups like "Temporarily laid off" and

"Looking for work" show medians which are near zero.

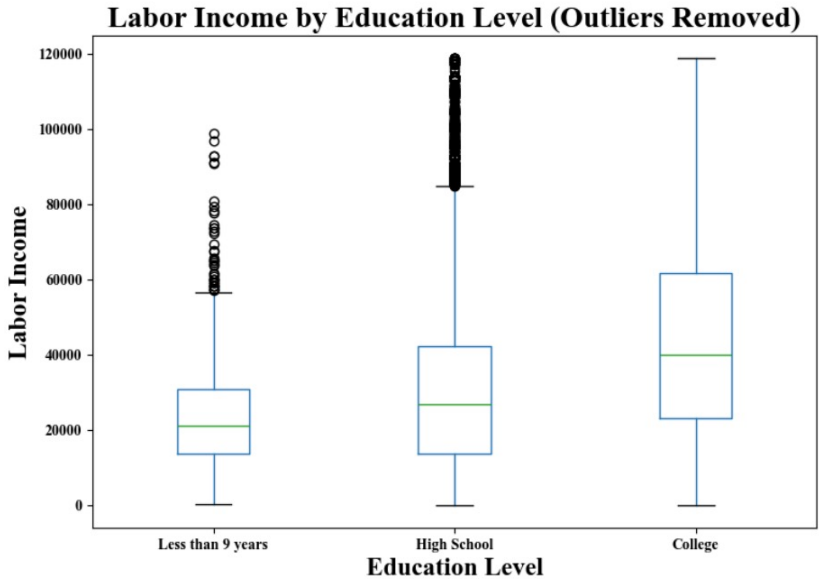

Figure 4: Labor Income by Education Level

Figure 4 is the boxplot of labor income versus education level. It is clear that people with higher education generally will have higher income. They have a positive correlation, and this also leads to income inequality. And it also shows that higher education will have wider income spread. For college, the box is taller and the whiskers are longer, so it indicates that there is more variation in income among people with higher education.

Figure 5: Labor Income by Gender

In Figure 5, the spell for males is much wider than that for females showing greater variation in income for males. Also, the maximum for males is substantially higher than that for females. In addition, the median labor income for males is higher than that for females. There are more outliers for females than males, which may imply that females may also achieve high income.

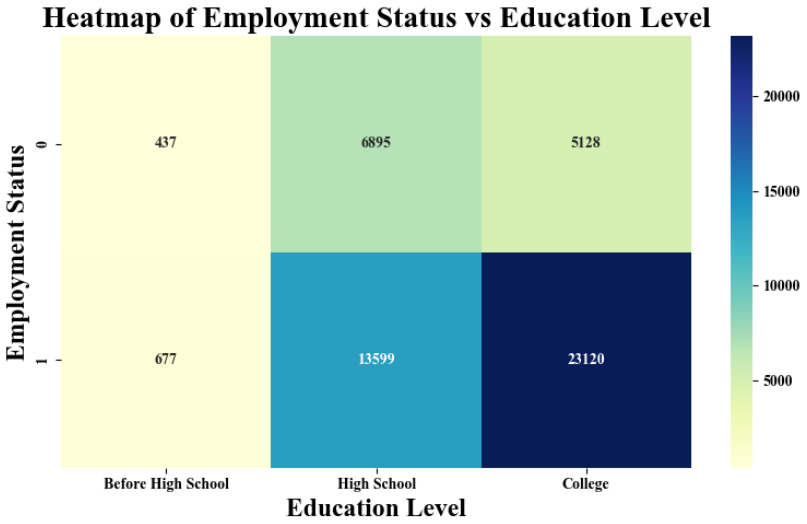

Figure 6: Years of Education vs Employment Status

From Figure 6, the darkest cell (highest count) is in the Employed column for college education (complete college or more) with 23,120 people. This suggests that most individuals achieve employment around this education level. Other notable employment peaks appear at the end of high school (9-12 years). The unemployment percentage is relatively high before high school education level, but drop off significantly after that. In addition, the percentage of people who are unemployed decreases when the education level increases. This indicates that higher education likely improves job stability.

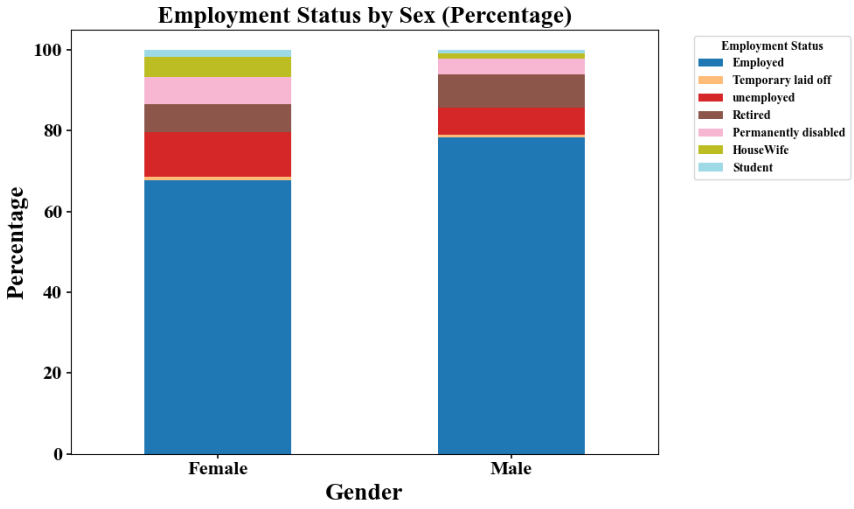

Figure 7: Sex vs Employment Status

For Figure 7, the largest portion of both groups is "Employed" (blue). However, males have a significantly higher employment rate than females. The Unemployed (red) and House-Wife (green) sections appear relatively larger for females. This can be related to the boxplot of the labor income difference between two genders, which implies that labor income may have some effects on employment status.

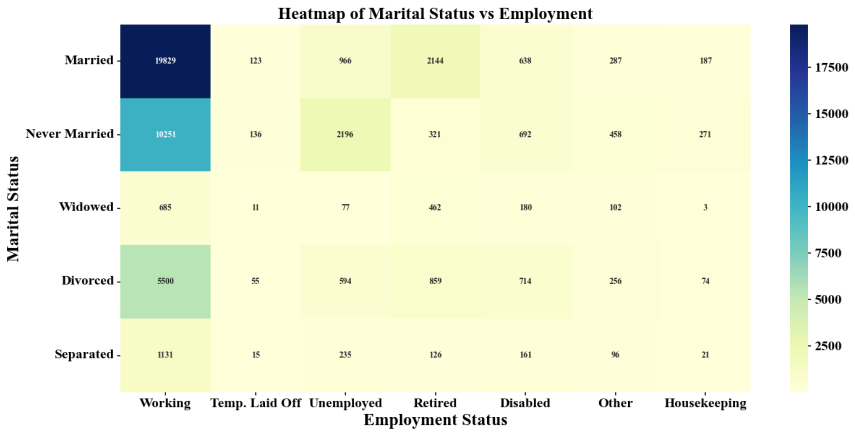

Figure 8: Marital status by employment

In Figure 8, the highest concentration of employed individuals (19829) are married. The heatmap suggests that employment is most common among married and never-married individuals. The unemployed category has notable numbers in the never married and divorced groups. This may suggest that divorce or lack of marriage correlate with higher unemployment.

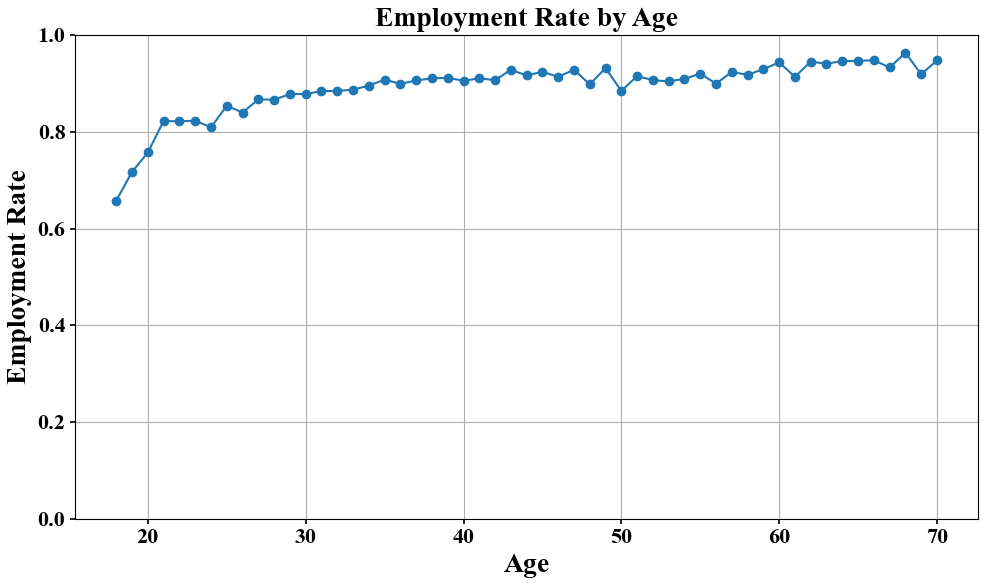

Figure 9: Employment rate by age

In Figure 9, the employment rate starts below 70 percent at age 18, then climbs quickly. Reflecting the transition from schooling to workforce entry. Also, the employment rate stabilizes at about 90 percent in the range 30 to 45. These are prime working age with highest labor force participation. Later, the employment rate has some up and downs. This situation may reflecting some health related exits and retirement of elder people. In addition, the employment rate is higher for older people may because of the exit from labor force. Elder people may experience age discrimination and retirement problems making them out of labor force. Then elder people who are currently in labor force will tend to have higher employment rate.

5 Methodology

Our approach to modeling employment dynamics is based on a two-stage logistic regression framework designed to reflect both initial labor market entry and long-run employment transitions. The objective was to simulate full employment trajectories for individuals from age 24 to 65. To achieve this, we first constructed two logistic models: one to predict initial employment status at age 24 (Model 1), and another to capture year-to-year employment changes for ages 25 and above (Model 2).

5.1 Full Models

Before the two-stage simulation structure, we began with a pair of comprehensive logistic regression models—one for initial probabilities and one for transition probabilities. Each incorporating the full set of available covariates. These full models served as an initial

benchmark to evaluate the predictive value of different features and to inform our ultimate model design.

5.1.1 Full Model 1: Employment Probability Across All Ages

Full Model 1 served as an initial baseline to explore how employment status correlates with a wide range of demographic and economic variables across all observed ages. This model was trained using logistic regression on the full dataset, spanning individuals aged 18 through 70.

To address class imbalance in the binary employment outcome (i.e., far more employed than unemployed individuals), we applied the SMOTETomek method, a hybrid resampling technique. SMOTE (Synthetic Minority Over-sampling Technique) generates synthetic examples for the minority class, while Tomek Links remove overlapping examples near class boundaries. Together, they help improve model balance and classification performance.

The covariates included age and its square, labor income, insurance coverage, education, gender, marital status, and survey year. We also constructed several interaction terms to capture more complex effects: specifically, interactions between education and gender, between age and income, and between labor income and year indicators. The following is

the equation:

$$\begin{aligned}
\log\left(\frac{P(\text{Employed}=1)}{1-P(\text{Employed}=1)}\right) = {} & 6.519 - 0.204 \cdot \text{Insured} - 4.342 \cdot \text{Edu: Some College} \\
& - 3.486 \cdot \text{Edu: College or Above} - 0.494 \cdot \text{Sex: Male} \\
& - 2.083 \cdot \text{Marital Status: Never Married} \\
& - 1.922 \cdot \text{Marital Status: Widowed} \\
& - 1.809 \cdot \text{Marital Status: Divorced} \\
& - 1.807 \cdot \text{Marital Status: Separated} \\
& - 1.879 \cdot \text{Year 2013} - 1.046 \cdot \text{Year 2015} \\
& - 0.800 \cdot \text{Year 2017} - 0.859 \cdot \text{Year 2019} - 1.154 \cdot \text{Year 2021} \\
& - 0.590 \cdot (\text{Edu: Some College} \times \text{Gender: Male}) \\
& - 1.087 \cdot (\text{Edu: College or Above} \times \text{Gender: Male}) \\
& - 0.023 \cdot \text{Age} + 0.0007 \cdot \text{Age}^2 + 0.0001 \cdot \text{Labor Income} \\
& - 2.158 \times 10^{-5} \cdot (\text{Labor Income} \times \text{Year 2021}) \\
& - 1.317 \times 10^{-5} \cdot (\text{Labor Income} \times \text{Year 2019}) \\
& - 3.743 \times 10^{-5} \cdot (\text{Labor Income} \times \text{Year 2017}) \\
& - 2.330 \times 10^{-5} \cdot (\text{Labor Income} \times \text{Year 2015}) \\
& - 1.475 \times 10^{-6} \cdot (\text{Age} \times \text{Income})
\end{aligned}$$

While Full Model 1 offered a broad cross-sectional view of factors associated with employment, it did not account for state dependence. As a result, the model was less suitable

for sequential employment modeling.

5.1.2 Full Model 2: Dynamic Employment with State Dependence

Full Model 2 extended the structure of Full Model 1 by incorporating a dynamic element: employment status in the previous period. This lagged employment variable was introduced to model persistence and transitions in employment, making it more appropriate for sequential simulation.

All other features remained consistent with Full Model 1. However, rather than using the entire dataset, we restricted the training sample to individuals with at least two consecutive survey years. The dataset consists of 28,285 observations from 9,258 individuals. The lagged employment indicator was one-hot encoded and included as an additional predictor. SMOTETomek oversampling was applied during training to handle class imbalance. The

following is the equation:

$$\begin{aligned}
\log\left(\frac{P(\text{Employed}=1)}{1-P(\text{Employed}=1)}\right) &= 4.204 - 0.227 \cdot \text{Insured} - 4.660 \cdot \text{Some College} \\
&\quad - 4.095 \cdot \text{College or Above} \\
&\quad + 0.046 \cdot \text{Male} + 0.880 \cdot \text{Employed Last Year} \\
&\quad - 2.157 \cdot \text{Never Married} - 2.450 \cdot \text{Widowed} - 1.812 \cdot \text{Divorced} \\
&\quad - 2.047 \cdot \text{Separated} \\
&\quad - 1.013 \cdot \text{Year 2013} - 0.574 \cdot \text{Year 2017} - 0.513 \cdot \text{Year 2019} \\
&\quad - 1.133 \cdot \text{Year 2021} \\
&\quad - 1.335 \cdot (\text{Some College} \times \text{Male}) - 1.717 \cdot (\text{College or Above} \times \text{Male}) \\
&\quad + 0.075 \cdot \text{Age} - 0.00005 \cdot \text{Age}^2 + 0.0001 \cdot \text{Labor Income} \\
&\quad - 3.682 \times 10^{-5} \cdot (\text{Labor Income} \times \text{Year 2021}) \\
&\quad - 4.305 \times 10^{-5} \cdot (\text{Labor Income} \times \text{Year 2019}) \\
&\quad - 5.494 \times 10^{-5} \cdot (\text{Labor Income} \times \text{Year 2017}) \\
&\quad - 3.435 \times 10^{-5} \cdot (\text{Labor Income} \times \text{Year 2015}) \\
&\quad - 1.275 \times 10^{-6} \cdot (\text{Age} \times \text{Labor Income})
\end{aligned}$$

The inclusion of employment lag significantly improved the model's ability to differentiate between persistent employment, short-term unemployment spells, and re-employment. In practice, Full Model 2 served as an early prototype of the reduced transition model later used in simulation. It provided valuable insight into how employment lag and other variables

interact over time.

5.2 Reduced Models for Simulation

To support simulation of employment trajectories from early to late career, we developed two reduced logistic regression models for longitudinal transition modeling. These models were intentionally simplified versions of the full specifications to ensure compatibility with the variables available in our client's proprietary datasets. This reduction in complexity enabled the simulation of employment dynamics for hypothetical individuals based solely on covariates the client can feasibly observe or impute, primarily education, gender, and labor income.

5.2.1 Reduced Model 1

Reduced Model 1 focuses exclusively on predicting whether an individual is employed at age 24—the starting point for simulation. The model is trained only on individuals who are 24 years old in the PSID dataset, thus avoiding age as a covariate and treating this prediction as cross-sectional. To address imbalance in employment status at entry, we also applied the SMOTETomek technique. The final model was fit using logistic regression and provided

robust predictions of initial labor force status. The following is the eqaution:

$$\begin{aligned}\text{logit}(\Pr(\text{Employed} = 1)) = \; & -0.0984 - 0.1093 \cdot \text{Edu: Some College} - 0.0357 \cdot \text{Edu: College or Above} \\ & - 0.0754 \cdot \text{Gender: Male} - 0.0676 \cdot (\text{Edu: Some College} \times \text{Male}) \\ & - 0.0253 \cdot (\text{Edu: College or Above} \times \text{Male}) \\ & + 5.1294 \times 10^{-5} \cdot \text{Labor Income}\end{aligned}$$

5.2.2 Reduced Model 2

Continuing to build on Reduced Model 1, we developed Reduced Model 2 to better capture employment dynamics over time. The primary goal of this model was to estimate the probability of transitioning into or remaining in employment, incorporating individuals' historical employment status. In addition to the variables used in Model 1, Model 2 introduced the lagged employment status to account for prior job outcomes. This lag captures whether an individual was employed in the previous year, allowing the model to reflect state dependence

and inertia in labor market outcomes. The following is the equation:

$$\begin{aligned}
\text{logit}(\Pr(\text{Employed} = 1)) = {} & 1.5514 \times 10^{-7} + 0.0001 \cdot \text{Edu: Some College} \\
& + 2.7988 \times 10^{-9} \cdot \text{Edu: College or Above} \\
& + 1.4609 \times 10^{-7} \cdot \text{Gender: Male} + 1.1638 \times 10^{-7} \cdot \text{Employed}_{t-1} \\
& + 3.2704 \times 10^{-7} \cdot (\text{Edu: Some College} \times \text{Gender: Male}) \\
& + 1.4810 \times 10^{-8} \cdot (\text{Edu: College or Above} \times \text{Gender: Male}) \\
& + 7.6703 \times 10^{-6} \cdot \text{Age} + 0.0004 \cdot \text{Age}^2 + 0.0001 \cdot \text{Labor Income} \\
& - 1.1537 \times 10^{-6} \cdot (\text{Age} \times \text{Labor Income})
\end{aligned}$$

5.3 Results

Before presenting model-specific results, we first outline the evaluation metrics used to assess model performance. Accuracy measures the overall proportion of correct predictions, but can be inflated when the majority class dominates. To address this, we include the weighted average F1-score, which combines both precision and recall, and adjusts for class imbalance by weighting each class according to its frequency. This makes the F1-score particularly useful for evaluating rare outcomes such as unemployment. Additionally, we report the Area Under the Receiver Operating Characteristic Curve (AUC), which captures the model's ability to rank individuals correctly based on their predicted probability of unemployment. An AUC of 1.0 indicates perfect discrimination, while a value of 0.5 suggests no better than random chance. Finally, we include pseudo R^2, which provides a general sense of how well the model fits the data. These metrics together provide a comprehensive view of each model's

classification performance, predictive robustness, and explanatory power.

Table 1: Model Evaluation Metrics

Model	**Accuracy**	**Weighted F1-score**	**AUC**	**Pseudo R^2**
Full Model 1	0.82	0.842	0.79	0.530
Full Model 2	0.85	0.870	0.80	0.573
Reduced Model 1	0.83	0.900	0.80	–
Reduced Model 2	0.83	0.954	0.84	–

As shown in Table 1, it summarizes key evaluation metrics—accuracy, F1-score, AUC, and pseudo R^2 —for all four logistic regression models.

Full Model 1 achieved a solid performance with 82% accuracy and an AUC of 0.79. Despite the relatively broad feature set, its pseudo R^2 of 0.53 suggests the model captures moderate explanatory power. The weighted F1-score of 0.842 indicates reasonable performance under label imbalance. Full Model 2 performed slightly better across all metrics after incorporating the lagged employment variable. Accuracy rose to 85%, AUC improved to 0.80, and the pseudo R^2 reached 0.573. These gains confirm the strong predictive value of state dependence in employment transitions. Individuals' prior labor market status is a critical determinant of future outcomes.

Reduced Model 1, trained only on age-24 individuals, still achieved 83% accuracy and an F1-score of 0.90, indicating a well-fitted model for predicting initial employment status with fewer variables. Its AUC matched Full Model 1, suggesting strong discriminatory power even with limited input features. Reduced Model 2 also performed competitively. The close alignment in results between full modes and reduced models further suggests that dimensionality reduction does not significantly compromise forecasting quality.

6 Simulation

After model estimation, we embedded both models into a Monte Carlo simulation framework. Each simulated individual begins their career at age 24 with characteristics sampled to match the observed distributions in the data. Their employment status is first determined by Model 1, and subsequent transitions are generated using Model 2, recursively updating features like age and lagged employment. This simulation allows us to study both average trends and heterogeneity in employment patterns over the life cycle.

6.1 Modeling Structure

Our simulation is driven by two logistic regression models:

- **Model 1** predicts the initial employment status at age 24. It is estimated using a cross-sectional subset of the data restricted to individuals aged 24 and includes covariates such as education level, gender, and labor income.
- **Model 2** models employment transitions from age 25 to 65. It incorporates age, a quadratic term in age, gender, education, labor income, and lagged employment status. Interaction terms, including Education $\times$ Gender and Age $\times$ Labor Income, are also included to capture heterogeneity in transition probabilities.

Each individual enters the simulation at age 24 and exits at age 65, resulting in a career span of 42 years. For each year, the model predicts the probability of being employed, and a Bernoulli trial is used to simulate the binary employment outcome.

6.2 Sampling Covariates

To ensure that our simulated population reflects realistic labor force heterogeneity, we implement a structured covariate sampling procedure informed by empirical distributions observed in the original dataset.

6.2.1 Education Level

Each simulated individual is randomly assigned an education level based on the empirical proportions in the dataset. The education categories are defined as follows:

Education 1 (Baseline): Less than high school

Education 2: High school graduate

Education 3: College degree or above

The probability of drawing each education level corresponds to its relative frequency in the dataset. This ensures that the distribution of education in the simulated population mirrors that of the observed sample.

6.2.2 Gender

Gender is sampled as a binary variable, with the probability of being male set to the empirical share of males in the dataset. User could also set gender as all males or all females. The gender variable is encoded as 1 for male and 0 for female.

6.2.3 Labor Income

Labor income is conditionally sampled based on education level. For each education category combined with each gender, we fit a log-normal distribution to the empirical income data.

The parameters for each combination are estimated as follows:

1. Compute the sample mean μ and standard deviation σ of labor income for each education group.

2. Transform the parameters into the log-normal scale using:

$$\sigma_{\log}^2 = \log\left(1 + \frac{\sigma^2}{\mu^2}\right), \quad \mu_{\log} = \log(\mu) - \frac{1}{2}\sigma_{\log}^2$$

3. Draw labor income from LogNormal$(\mu_{\log}, \sigma_{\log})$.

This approach preserves the right-skewed distribution of income and captures heteroskedasticity across education levels. Simulated incomes are constrained to be non-negative and truncated below at zero.

6.2.4 Covariate Composition

For each individual, we store the following baseline covariates:

`Education_2`: indicator for high school graduate (vs. baseline)

`Education_3`: indicator for college graduate (vs. baseline)

`Gender_1`: indicator for male or not

`Labor_Income`: sampled income value

These covariates are then used to generate time-varying features such as age, age-squared, and interaction terms as the simulation progresses.

6.3 Simulation Logic

The simulation proceeds recursively for each individual:

1. At age 24, employment status is determined using Model 1 and the sampled covariates.

2. From age 25 onward, employment status is determined using Model 2. The model includes lagged employment status, updated age variables, and other static covariates.

3. For each year, we compute the log-odds of employment and transform it into a probability via the logistic function:

$$P(\text{Employment}) = \frac{1}{1 + \exp(-\text{log-odds})}$$

4. Gaussian noise is added to each probability to reflect unobserved heterogeneity and avoid overly deterministic trajectories.

5. A Bernoulli draw is used to simulate the binary employment outcome.

This process is repeated for each of the 10,000 individuals in the simulation.

6.4 Simulation Results and Conclusions

6.4.1 Unemployment Trends and Transition Probabilities

Figure 10 compares the simulated unemployment rate by age to the empirical unemployment rate in the dataset, alongside simulated transition probabilities $P(U_t \mid E_{t-1})$ and $P(U_t \mid U_{t-1})$. Several important patterns emerge:

To begin with, the simulated unemployment rate closely tracks the age trend seen in the real data, generally declining with age. However, it slightly overestimates unemployment in early years (ages 25–30) and underestimates it beyond age 55.

The conditional probability of becoming unemployed given employment in the previous year ($P(U_t \mid E_{t-1})$) remains low and stable across ages, around 10% on average.

In contrast, the probability of remaining unemployed conditional on previous-year unemployment ($P(U_t \mid U_{t-1})$) is significantly higher—typically between 15% and 25%—and exhibits more variability across the life cycle.

These results suggest that the model captures key features of labor market persistence: once unemployed, individuals are more likely to remain so. However, re-employment is reasonably likely across age groups, contributing to declining unemployment rates as individuals age.

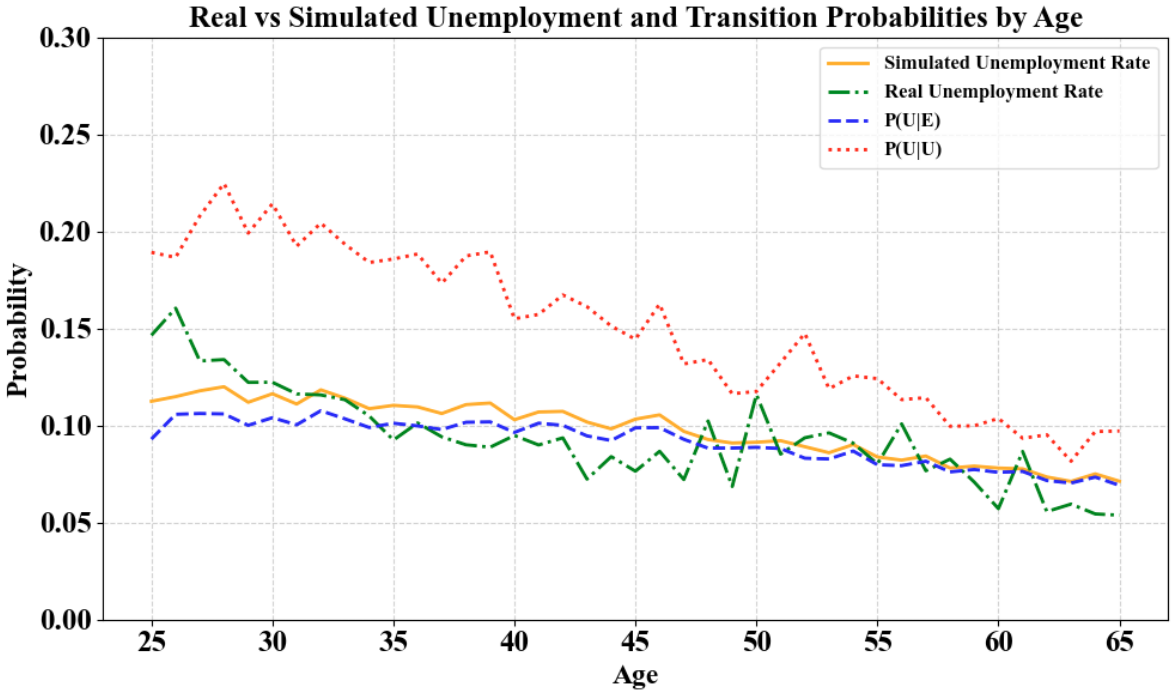

Figure 10: Real vs. Simulated Unemployment Rates and Transition Probabilities by Age

6.4.2 Unemployment Spell Lengths

Figure 11 displays the distribution of simulated unemployment spell lengths across all individuals. The distribution is highly right-skewed:

The majority of unemployment spells last only one year, with over 30,000 occurrences in the sample. Two- and three-year spells are significantly less frequent, and spells longer than four years are extremely rare.

This distribution reflects a transition model in which re-employment probabilities are high, particularly after one or two years of unemployment.

While the simulation successfully captures the prevalence of short-term unemployment, the near absence of long-term spells suggests that the current model may understate the persistence of unemployment for certain subgroups. Future extensions could incorporate duration dependence, macroeconomic shocks, or heterogeneous re-employment probabilities to generate a heavier right tail.

Figure 11: Unemployment Spell Length

6.4.3 Simulated Unemployment Rates by Gender

Knowing that our simulation model captures the overall dynamics of unemployment across the life cycle, we proceed to examine whether it also replicates key demographic disparities. One such disparity is the well-documented gender gap in labor market outcomes. To this end, we generate separate simulations for males and females, using gender-specific covariate distributions and applying the same transition logic governed by the estimated logistic regression models.

As illustrated in Figure 12, the simulated unemployment rates for females are consistently higher than those for males at every age from 24 to 65. This aligns with established labor market trends and highlights the model's ability to reflect structural heterogeneity embedded in the input features. Female individuals in the real dataset tend to have slightly lower

labor income and different education distributions, both of which influence the simulated employment trajectory through their roles in the predictive equations.

In summary, the gender-specific simulations not only reinforce the external validity of the model but also highlight its usefulness in studying inequality in labor market experiences. The fact that the simulated gender gap aligns with empirical evidence adds further credibility to the model's structure and parameterization.

Figure 12: Simulated Unemployment Rate by Gender

7 Recommendation

Based on our simulation results, we propose several ways in which the model outputs can be operationalized to inform investment strategy and product design. Our model identifies individuals who are at heightened risk of experiencing prolonged or recurring unemployment, often characterized by lower education levels, reduced labor income, or weaker employment

histories. These risk profiles can inform more tailored investment strategies. Specifically, high-risk individuals may benefit from more conservative asset allocations or larger liquidity buffers to mitigate the impact of career interruptions. By incorporating employment risk into portfolio design, financial planning can become more adaptive to individual labor market realities.

In addition to portfolio-level insights, simulation results can be used to guide the design and delivery of retirement-related products. Clients identified as vulnerable to income instability could be directed toward solutions that prioritize income smoothing, enhanced annuitization options, or longer accumulation periods. Communication strategies can also be adapted to encourage higher savings behavior among these groups or prepare them for longer working horizons.

8 Future Directions

Our simulation framework offers a strong foundation for studying employment dynamics, but there are multiple opportunities for refinement and expansion.

First, while this project primarily relied on logistic regression, future studies could explore alternative modeling approaches such as the probit model. The probit model extends to multinomial or ordinal settings with outcomes beyond the binary framework. For example, in addition to employment and unemployment, a third state such as retirement or withdrawal from the labor force could be modeled explicitly. Incorporating such multi-state transitions would enable richer and more realistic labor market dynamics to be captured, especially for older individuals approaching the end of their careers.

Second, the current simulation framework models employment dynamics using only a single one-period lag. Extending this to incorporate multiple lags—such as a two- or three-period employment history—would allow the model to capture longer-term persistence in labor force participation. Moreover, additional covariates could be allowed to evolve over time, including marital status, educational attainment, or household composition. While the current specification includes core demographic and economic indicators, more socioeconomic covariates could further improve the model's ability to capture heterogeneity in unemployment risk.

Finally, handling class imbalance remains a critical consideration. Although SMOTE-Tomek oversampling was used in training, future implementations may benefit from alternative approaches, such as adaptive synthetic sampling (ADASYN), cost-sensitive loss functions, or direct threshold adjustment based on simulation objectives to ensure better alignment between classification performance and simulation fidelity.

References

Panel Study of Income Dynamics (2025). *Panel Study of Income Dynamics (PSID)*. https://psidonline.isr.umich.edu. Accessed: May 11, 2025.